Published in: Šticha, František/Fried, Mirjam (eds.): Grammar & Corpora 2007. Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice - Prag: Academia, 2008. S. 251-260.

Grammatical variation in Near-Standard German: a corpus-based project at the Institute for the German Language (IDS) in Mannheim

Jacqueline Kubczak – Marek Konopka, Mannheim
kubczak@ids-mannheim.de, konopka@ids-mannheim.de

1 General description

'Grammatical variation in Near-Standard German' is a new project at the Grammar Department of the Institute for German Language (IDS) in Mannheim. The project comprises preparatory work for a corpus-based grammar of standard and close-to-standard German. It will start officially in the summer of 2008 but pilot studies have already begun. The new grammar itself will be written from 2013 onwards. In comparison with existing grammars of German, especially with the three-volume IDS-Grammar (Zifonun et al. 1997), it will introduce some important innovations. The corpus-based approach makes the following improvements possible:

- a detailed investigation of language use, focusing on variation, frequency, and distribution of grammatical features across text types/registers/varieties;
- greater attention to details aiming at a description of still unrecorded patterns; and
- greater reliability of the findings.

A comparable project for the English language led to the *Longman Grammar of Spoken and Written English* published in 1999 (Biber et al. 1999) and, at the current stage of technological development, a comprehensive corpus-based grammar of German has also become due. Traditional, introspective grammar and its disconnect from empirical research has often suggested a more or less uniform, barely differentiated standard language. However, due to sophisticated analyses of ever-growing databases, the usage of language can now be investigated in its full range of variation, and it can be reliably modelled with respect to parameters such as text type, register, and social and regional variety. As regards public interest, this approach is highly promising. In the fields of language proficiency testing, first language pedagogy, and teaching German as a second language, there is a high demand for solid understanding of the scope of available grammatical variants and the way of using them actively in contemporary language.
The project Grammatical variation within Near-Standard German utilises the corpus-linguistic infrastructure of the IDS. The project's main data source is Das

Deutsche Referenzkorpus (DeReKo – the Mannheim German Reference Corpus).[1] Currently, DeReKo contains over two billion words and allows the creation of virtual corpora, which can be differentiated according to specific requirements. The main tool for examining corpus data is the Corpus Search, Management, and Analysis System, COSMAS II.[2] Moreover, the project must include statistical analyses that are not driven by linguistic hypotheses. The research on such kinds of analyses is carried out by the IDS-corpus-linguistic group. It is treated in Keibel et al. in this volume.

So far, numerous problematic areas and variation fields, which are to be examined more thoroughly, have been identified in the preparatory work carried out by the research teams Grammatik in Fragen und Antworten ('Grammar of questions and answers') and Konnektoren ('Connectors'). The 'Grammar of questions and answers' team has collected over 250 borderline cases and specific difficulties of German grammar that are mostly based on variation.[3] At present, the team is working on the usage-oriented description of these problematic cases in the framework of the Internet Grammar Information System Grammis.[4] Some examples of the difficulties treated are:

- variation of preterite and participle verb forms (e.g., *buk/backte, gewinkt/gewunken*);
- variation within the inflection of attributive adjectives (e.g., *einige interessante(n) Themen*);
- variation in case marking after prepositions (e.g., *wegen des/dem Wetter*);
- use of infinitival clauses (the positioning of the infinitival clause and the use of *zu*); and
- word order variation after specific connectors.

As regards the realisation of the project Grammatical variation within Near-Standard German, there are three major tasks:

- **The development of a feasible concept of near-standard variation**: It could, for example, include the variation within genres or text types that are destined to be understandable across social and regional varieties and are expected to omit socially or locally limited grammatical features. However, the concept should still include variation among regional varieties of a higher level such as the national varieties of the Federal Republic of Germany, Switzerland, and Austria.
- **The design of a corpus** representing the core types of text and discourse in German in an adequate quantitative relation: the corpus must include also

[1] Cf. http://www.ids-mannheim.de/kl/projekte/korpora/.
[2] Cf. http://www.ids-mannheim.de/cosmas2/.
[3] Cf. http://www.ids-mannheim.de/grammatikfragen/.
[4] Cf. http://www.ids-mannheim.de/grammis/.

those areas of language in which one can expect deviations from the grammatical norm. Colloquial language, substandard usage in the transition between spoken and written language as used in e-mails and internet forums, and also texts by non-native speakers with a higher level of proficiency are taken into account. Furthermore, a subcorpus of spoken language could also be included. The whole corpus should be morpho-syntactically annotated. The analyses enabled by annotations and the grammar employed by the analytical tools are intended to be as empirical and theory-neutral as possible. Such an approach allows the recognition of new, exceptional, and non-standardised structures and language patterns. So far, only a small part of DeReKo (containing 26 million words) can be regularly analysed by means of morpho-syntactic categories. Such queries with the entire DeReKo are currently being tested but they are not yet available with the aid of COSMAS II.

- **The completion of four studies** related to phonology, morphology, syntax, and text: these studies should exemplify one variation field per linguistic domain. Possible issues are, for example, variation in the use of infinitival constructions in the syntax domain and variation in the use of connectors and text organization markers in the text domain. The descriptive IDS-grammar (Zifonun et al. 1997) serves as the theoretical basis for the studies.

In the following short preview, a small extract of our research work on infinitives with and without *zu* will be presented. This variation field is a notorious problem in terms of usage and linguistic description. In another paper in this volume, Ulrich H. Waßner addresses the variation in the use of connectors within the framework of the same project.

2 Subject infinitival clauses with and without *zu*

2.1 Introduction

The IDS-project 'Grammar of questions and answers" already tries to meet the demands of language users by describing contemporary language use instead of presenting prescriptive rules. Therefore, the DeReKo corpus is the basis for the research. One of the frequently asked questions is about infinitival clauses governed by a verb. In German, such infinitival clauses may occur with or without the particle *zu*. This often leads to uncertainties about appropriate usage.

In order to remain concise, the focus of this paper will be restricted only to subject infinitival clauses and even more precisely, to preverbal subject infinitival clauses, since subject infinitival clauses following the main clause verb always need *zu*, whereas preverbal subject infinitival clauses can occur with or without *zu*.

2.2 Preverbal subject infinitival clauses

The variable usage of *zu* in subject infinitival clauses can be illustrated by the following corpus examples:

(1) [INF+*zu*] *Dieses Buch* ***zu*** *lesen strengt an, weil es Gefühle aufwühlt und Assoziationsketten durchs Hirn jagt.* (Berliner Zeitung, 02.12.2002, p. 13)
'To read/Reading this book is strenuous because ...'

(2) [INF-*zu*] *Immer progressiv* ***[Ø]*** *sein, strengt eben an.* (TAZ, 11.04.1992, p. 28)
'To be progressive all the time is just strenuous.'

2.3 Frequency of the phenomenon

First, with the help of the IDS Research Programme for Corpus Linguistics (especially Cyril Belica and Marc Kupietz), a list of verbs with preverbal subject infinitival clauses was compiled in order to examine the principles of its variable usage. Then, the occurrence of infinitives with and without *zu* with all of these verbs was investigated. The first result of this analysis is the generalization that preverbal subject infinitival clauses are more frequent in German than in English. They occur approximately 120 times per million words in German but fewer than 50 times per million words in English (cf. Biber et al. 1999).

Table 1 DeReKo analysis based on morpho-syntactic annotation with the aid of the Stuttgart Tree-Tagger[5] (Schmid 1995)

	sentences
DeReKo corpus size |C|	122040690
pattern hits |PH|	642905
sample size |S|	2999
true pattern hits in sample |H|	1013
false pattern hits in sample	1986
pattern/tagger accuracy	33.78%
estimated occurrences in C	217159
estimated frequency of phenomenon	1779.41 per million sentences ≈ 120 per million words

2.4 The parameters influencing the choice between INF+*zu* and bare INF

The analysis of the language material led to a hypothesis about parameters that might affect the choice between infinitival clauses with *zu* and those without *zu*. This hypothesis is to be verified by means of statistical analyses on an extensive corpus. The parameters are:

[5] The complete regular expression applied to the tagger is quoted in the appendix.

- main clause verb,
- length of the infinitival clause,
- definiteness or indefiniteness of the complements of the infinitive,
- tense and voice of the infinitival clause, and
- perhaps others.

2.4.1 The main-clause verb

The IDS-Grammar (Zifonun 1997) states that *zu* is optional in preverbal subject infinitival clauses; that is, it is not determined by the main-clause verb. In theory, this is true but the aim of the project is to investigate the actual usage and, more precisely, to find out whether certain verbs are more likely to be used with one of the two constructions even if the choice of the construction is unrestricted in principle. The results are integrated in Table 2:

Table 2 Main-clause verbs and preverbal subject infinitival clauses (an extract: 20 most frequent verbs)

VFIN lemma	**hits**	%	**cum. %**	**est. # in C**	**% with *zu***	**est. # in C with *zu***	**est. # in C without *zu***	**(ø # words before VINF with *zu*) – (ø # words before VINF without *zu*)**
sein	328	32.38	32.38	70314	93.60	65812	4501	2.75
heißen	115	11.35	43.73	24652	43.48	10718	13934	1.46
werden	61	6.02	49.75	13076	95.08	12433	643	2.02
bedeuten	50	4.94	54.69	10718	70.00	7503	3215	2.00
machen	37	3.65	58.34	7931	72.97	5788	2143	2.65
haben	25	2.47	60.81	5359	84.00	4501	857	4.24
bringen	23	2.27	63.08	4930	56.52	2786	2143	3.15
kommen	21	2.07	65.15	4501	71.43	3215	1286	3.87
scheinen	21	2.07	67.23	4501	95.24	4287	214	N/A
fallen | fällen	16	1.58	68.81	3429	75.00	2572	857	0.92
gehören	16	1.58	70.38	3429	87.50	3001	428	N/A
bleiben	14	1.38	71.77	3001	85.71	2572	428	N/A
reichen	13	1.28	73.05	2786	53.85	1500	1286	0.45
gehen	13	1.28	74.33	2786	69.23	1929	857	4.39
lassen	11	1.09	75.42	2358	54.55	1286	1071	5.20
kosten	11	1.09	76.51	2358	90.91	2143	214	N/A
lohnen	10	0.99	77.49	2143	60.00	1286	857	4.25
gelingen	10	0.99	78.48	2143	100.00	2143	0	N/A
gelten	10	0.99	79.47	2143	70.00	1500	643	5.10
erscheinen	9	0.89	80.36	1929	100.00	1929	0	N/A

The table shows that:

a) Many verbs can be used with a preverbal subject infinitival clause (467 verbs are attested in the corpus) but for most of these verbs, the preverbal subject infinitival clause is only a peripheral phenomenon. Over 80% of the findings are covered by only 20 verbs and over 50% are covered by only 4 verbs (*sein, heißen, warden,* and *bedeuten*). Such information can be important for teaching German as a Second Language, applied linguistics, and other areas.
b) None of the 20 verbs occurs exclusively with infinitival clauses of one or the other kind.[6]
c) Infinitival clauses with *zu* are clearly predominant. This means that they can be considered to be the norm.
d) The majority of the infinitival clauses without *zu* co-occur with the verbs *heißen, reichen, lassen, bringen, lohnen, gehen, bedeuten,* and *gelten.* This phenomenon can also be important for teaching German as a Second Language.
e) The verbs *kosten, sein, werden, scheinen,* and – most of all – the verbs *gelingen* and *erscheinen* occur almost exclusively with infinitival clauses with *zu.*

All of this allows us to draw the conclusion that there is a kind of usage-oriented verbal government.

2.4.2 The length of the infinitival clause

The corpus investigation revealed that a high number of infinitives without *zu* consists only of the infinitive (without complements). A comparison of the average length of the infinitival clauses with and without *zu* shows significant differences between the two (cf. the last column in Table 2). One can, therefore, assume that there is a tendency for *zu* to occur in longer subject infinitival clauses, whereas infinitives without *zu* are more easily found in shorter constructions.

2.4.3 Definiteness or indefiniteness of the complements of the infinitive

The corpus investigation also revealed that many of the infinitival clauses without *zu* are based on monovalent verbs such as *lachen* or *sterben.* Polyvalent verbs, however, can also be used without complements when the action they refer to is cast as a generalisation, cf. *geben ist schöner als nehmen* ('to give is better than to receive'). Generalizations can also be indicated by the use of nouns with an indefinite article in the singular or without an article in the plural form. If generalization proves to be a reason for using an infinitive without *zu,* there could be a difference in the occurrence of definite articles in infinitival clauses with and

[6] The occurrences of the two verbs quoted in the table with 100% infinitival clauses with *zu* (*erscheinen* and *gelingen*) have been verified in the entire DeReKo (and not only in the 3000-sample): there were 98% occurrences of INF+*zu* with *erscheinen* and 97% with *gelingen.*

without *zu*. Our first pass through the data indicates that INF-*zu* clauses with a definite extension are very rare and so it really seems to be more difficult to use an infinitive without *zu* when it occurs with a definite complement. An example is in (3):

(3) a. *Ein gutes Buch / Gute Bücher lesen strengt an.*
b. *Das gute Buch / Diese guten Bücher zu lesen, strengt an.*
c. (?) *Das gute Buch / Diese guten Bücher lesen, strengt an.*

For the moment, this is more or less an assumption and the relevance of the parameter definiteness or indefiniteness needs to be statistically validated.

2.4.4 Tense and voice of the infinitival clause

The corpus examination showed that preverbal subject infinitival clauses in the past tense or passive voice occur nearly exclusively with *zu*. This was checked with the verbs *heißen* and *bedeuten*. Only one sentence without *zu* in the past tense and three in the passive voice could be found. The most common uses are in (4), rare occurrences are in (5):

(4) past tense: *In London gelebt **zu** haben, heißt in gewissen Kreisen, die Alpes d'Huez des Studentenlebens geschafft zu haben.* [Berliner Zeitung, 05.01.2005]
'To have lived in London,...'
passive: *Dort aufgenommen **zu** werden, bedeutet eine Doppelbelastung.* [TAZ, 06.07.2004]
'To be affiliated there, ...'

(5) past tense: *Vieles erfahren haben, heißt noch nicht Erfahrung besitzen.* [Tiroler Zeitung, 12.01.2005, p. 27]
‚To have experienced a lot, ...'
passive: *Von Sicherheit eingewiegt werden, bedeutet sicherer* (sic) *Tod.* [Mannheimer Morgen, 02.10.1999]
'To be lulled by safety, ...'

These observations must also be validated by further investigation.

2.4.5 Additional parameters?

There are certainly other parameters that should be analysed, such as, for example, the position of the main verb in the sentence, but besides grammatical parameters of this kind, there also seem to be other factors influencing the choice between an infinitive with and without *zu*.

The most striking verb of the entire analysis is *heißen*. Its 57% infinitival clauses without *zu* could surely be partly related to traditional usage of, and variation in, popular sayings such as *Alles verstehen heißt alles verzeihen* and *Von der Sowjetunion lernen heißt siegen lernen*. In fact, the two sayings and their variants are copiously represented in the findings for the verb *heißen*.

2.5 Conclusion

By means of statistical corpus analyses, it was possible to establish preferences of individual verbs for preverbal subject infinitival clauses with or without *zu* and to verify that infinitival clauses without *zu* are normally shorter than those with *zu*. In relation to this, the relevance of factors such as definiteness of the infinitival complements, or the tense and voice of the infinitival clause could be pointed out. The significance of such factors, however, needs to be statistically substantiated. We are still at the very beginning of our efforts and there is much to be done. A reliable comparison of varieties such as spoken and written German, different genres, and dialects remains still a desideratum, and so does the improvement of the morpho-syntactic annotation in order to increase the amount of positive results.

Acknowledgements
We are grateful to Gisela Zifonun, who developed the work plan for the Grammar Department of the IDS, and Matthias Moesch, who helped us to prepare the talk.

Appendix

The complete (perl) regular expression looks like this:

```
/^ (WORD=(?!,)[^ ]* CLASS=(?!(?:PTKZU|V.FIN))[^ ]*
LEMMA=(?!(?:daß|dass|ob|weil|statt|um|obwohl|nachdem|was|wem|wer|
wen|wann|wenn|obgleich|obschon|solange|indem|sobald))[^ ]* )*(WORD=zu
CLASS=PTKZU LEMMA=zu )?(WORD=[^ ]* CLASS=V.INF LEMMA=[^ ]* )
(WORD=, CLASS=[^ ]* LEMMA=[^ ]* )?(WORD=(?:(?=[^ ]*schien )|(?!(?:sind
|[^ ]*st |[^ ]*en |(?:gab|gibt) CLASS=V.FIN LEMMA=geben WORD=es )))[^ ]*
CLASS=V[VA]FIN LEMMA=(?!UNKNOWN)[^ ]* )/
```

It was applied to the TreeTagger output, where each sentence was combined in a single line of the following form:

```
WORD=Verkrampfen CLASS=VVINF LEMMA=verkrampfen WORD=heißt
CLASS=VVFIN LEMMA=heißen WORD=verlieren CLASS=VVINF
LEMMA=verlieren WORD=. CLASS=$. LEMMA=.
```

For clarity and to avoid errors, the regular expression was assembled as follows:

```
my $ANY_WORD                = "WORD=[^ ]*";
my $ANY_CLASS               = "CLASS=[^ ]*";
my $ANY_LEMMA               = "LEMMA=[^ ]*";
my $ANY_KNOWN_LEMMA = "LEMMA=(?!UNKNOWN)[^ ]*";

my $LCTX_WORD               = "(?!,)[^ ]*";
my $LCTX_CLASS              = "(?!(?:PTKZU|V.FIN))[^ ]*";
my $LCTX_LEMMA              = "(?!(?:daß|dass|ob|weil|statt|um|obwohl|nac
hdem|was|wem|wer|wen|wann|wenn|obgleich|obschon|solange|indem|so
bald))[^ ]*";

my $ACCEPTED_LCTX           = "WORD=$LCTX_WORD CLASS=$LCTX_
CLASS LEMMA=$LCTX_LEMMA ";
my $PTK_ZU                  = "WORD=zu CLASS=PTKZU LEMMA=zu ";
my $ANY_INFIN               = „$ANY_WORD CLASS=V.INF $ANY_LEMMA
„;
my $A_COMMA                 = „WORD=, $ANY_CLASS $ANY_LEMMA „;
my $V_A_FINV                = „$ANY_WORD CLASS=V[VA]FIN $ANY_
LEMMA „;
my $V_A_FINV_3SG            = „WORD=(?:(?=[^ ]*schien )|(?!(?:sind |[^ ]*st |[^
]*en |(?:gab|gibt) CLASS=V.FIN LEMMA=geben WORD=es )))[^ ]* CLASS=V[VA]
FIN $ANY_KNOWN_LEMMA „;

my $PATTERN     = „^
($ACCEPTED_LCTX)*($PTK_ZU)?($ANY_INFIN)($A_COMMA)?($V_A_
FINV_3SG)";
```

Possible sources of error:
Both the TreeTagger and the regular expression that filters out candidate sentences with infinitives as the subject may have introduced a type 2 error concerning these candidates that may have biased the results. This has to be investigated further.

References

BIBER, D. et al. (1999): *Longman grammar of spoken and written English.* London: Longman.
MANN, H. B. – WHITNEY, D. R. (1947): On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18, 50–60.
SCHMID, H. (1994): Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of International Conference on New Methods in Language Processing,* Manchester, UK, 44–49.
SCHMID, H. (1995): Improvements in part-of-speech tagging with an application to German. In: Feldweg, H. – Hinrichs, E. (eds.), *Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen,* Tübingen: Niemeyer, 47–50.
WELCH, B. L. (1947): The generalization of 'student's' problem when several different population variances are involved. *Biometrika* 34, 28–35.
ZIFONUN, G. et al. (1997): *Grammatik der deutschen Sprache.* Berlin/New York: de Gruyter.